creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

보건학석사 학위논문

지역사회건강조사에서 이상조사원의 탐색 및 영향 분석

Analysis and exploration of falsified interviewers in Korean Community Health Survey (KCHS)

2015년 2월

서울대학교 보건대학원

보건학과 보건통계전공

원 은 지

지역사회건강조사에서

이상조사원의 탐색 및 영향 분석

지도교수 김 호

이 논문을 보건학 석사 학위논문으로 제출함

2014년 12월

서울대학교 보건대학원

보건학과 보건통계학전공

원은지

원은지의 보건학 석사 학위논문을 인준함

2014년 10월

위 원 장 조성일 (인)

부 위 원 장 원성호 (인)

위 원 김 호 (인)

국문초록

지역사회건강조사에서 이상조사원의 탐색 및 영향 분석

원은지(Eunji Won)

보건학과 보건통계전공(Dept. of Epidemiology and Biostatistics)

서울대학교 보건대학원

연구배경

지역사회건강조사는 시·군·구 단위의 주민 건강통계를 생산하기 위한 우리나라 최초의 조사 사업으로 질병관리본부와 광역자치단체 및 전국 253개 시·군·구 보건소가 역할을 분담하여 지역단위 건강조사를 효율적으로 수행해오고 있다. 다양한 종류의 수행기관과 인력이 동시에 투입되기 때문에 자료 생산의 정확성과 신뢰성을 위해서는 조사과정 전반에 걸친 질 관리(Quality Control)가 무엇보다도 중요하다. 지역사회건강조사와 같이 1:1 면접(face-to-face)을 통한 설문조사는 자료의 질적인 부분에서 조사원이 미치는 영향이 크다. 조사원이 의도적으로 본래의 조사 규정과는 다르게 조사하는 경우 위조 행위 조사(interviewer falsification)또는 부정행위 조사(interviewer cheating)로 정의하고 부정 행위를 하는 조사원(cheating interviewer)은 정상적으로 조사를 완료한 조사원(honest

interviewer)과 비교하여 어떤 특성을 가지고 있고 설문 조사에 어떻게 영향을 미치는지 연구한다.

목표

본 연구 목적은 부정행위를 한 조사원(cheating interviewer)의 설문조사 데이터와 부정행위를 하지 않고 폐기된 자료가 없이 정상적으로 완료한 조사원(honest interviewer)의 설문조사 데이터를 비교하여 부정행위를 한 조사원(cheating interviewer)의 특성을 찾는다. 또한 이러한 특성을 통하여 부정행위를 할 가능성이 있는 의심 조사원(at risk interviewer)을 탐색할 수 있는 방법을 마련하여 질 관리(Quality Control) 향상을 통해 지역사회건강조사의 정확성과 신뢰성을 확보하는 것이다.

연구방법

본 연구는 질병관리본부 주관으로 시행된 2012년 지역사회건강조사 원시자료(raw data), 2012년 지역사회건강조사 폐기자료, 2012년 지역사회건강조사 조사원 질관리 지표 자료, 2012년 지역사회건강조사 조사원 특성 자료를 이용하였다. 기존에 연구된 방법과 동일하게 기타 응답율(other-answers-ratio), 극단적 응답율(extreme-answers-ratio), 응답거부율(non-answers-ratio), Benford 분포를 이용한 x^2 값, 필터 문항 응답율(filter-answers-answers-ratio)의 총 다섯 가지 지표(indicators)로 클러스터 분석(cluster analysis)을 하여 부정행위를 한 조사원의 특성을 탐색하였다.

결과

대리 응답으로 자료가 폐기된 조사원 10명을 '부정행위를 한 조사원(cheating interviewer)' 그룹으로 분류하고, 부정행위를 하지 않고 폐기된 자료가 없으며 전화 점검의 재조사(re-interview)를 통해 응답일치도가 6인 조사원 12명을 '정상 조사원(honest interviewer)' 그룹으로 분류하였다. 두 그룹에서 다섯 가지의 지표를 비교했

을 때, 부정행위를 한 조사원(cheating interviewer) 그룹의 기타응답비율(other-answers-ratio)과 극단적 응답율(extreme-answers-ratio)의 값이 가정과 일치하게 낮았고 부정행위를 한 조사원(cheating interviewer) 그룹의 Benford 분포를 이용한 x^2 값과 필터문항 응답율(filter-answers-answers-ratio)은 가정과 일치하게 높았다. 반면에 응답거부율(non-answers-ratio)은 가정과 반대의 결과가 나왔다. 뿐만 아니라 다섯 가지 지표(indicators)을 이용하여 Ward Hierarchical Cluster 와 K-Means Cluster 분석을 한 결과 부정행위를 한 조사원(cheating interviewer)을 탐색하는 확률이 각각 60%, 90%가 나왔다. 또한, 폐기된 자료의 전체 조사원 65명(조사건수:256)과 응답일치도 6인 전체 조사원 18명(조사건수:288)의 질 관리 지표를 비교했을 때 조사 소요시간 이상치 비율(9분 이내)에서 차이가 컸고 나머지 질 관리 지표값은 큰 차이가 없거나 폐기된 자료의 조사원이 더 우수했다.

결론

설문 조사는 부정행위를 하는 조사원(cheating interviewer)에 의해 잠재적으로 영향을 받게 되고 조사 중 부정행위는 심각한 bias를 야기시키는 것으로 밝혀졌다. 현존하는 많은 방법들이 부정행위를 하는 조사원을 추출하기 위해 재조사(re-interview)를 시행한다. 재조사(re-interview)는 부정행위를 하는 조사원(cheating interviewer)을 찾는 가장 흔한 방법이지만 비용이 많이 들기 때문에 조사한 전체 대상자에 대해 재조사(re-interview)를 하는 것은 사실상 불가능하다. 그렇기 때문에 부정행위를 할 가능성이 있는 조사원(at risk interviewer)을 탐색하여 재조사(re-interview)를 하는 것이 가장 이상적이다. 현재 전체 22만 명에서 랜덤으로 10%를 추출하여 전화 점검으로 재조사를 시행하고 있지만 앞에서 분석한 다섯 가지의 지표로 다변량 분석을 통해 부정행위를 할 가능성이 있는 조사원(at risk interviewer)을 탐색 하여 우선적으로 재조사를 시행할 수 있다. 또한, 폐기된 자료의 전체 조사

원은 응답일치도가 6인 정상적으로 완료한 조사원에 비해 조사시간 이상치 비율이 컸다. 부정행위를 할 가능성이 있는 조사원(at risk interviewer)을 탐색하기 위해서 조사시간에 대한 심층적인 분석으로 지역사회건강조사의 질 관리를 향상시킬 수 있을 것이다.

주요어: 부정행위를 한 조사원, 지역사회건강조사, 질 관리, 벤포드 법칙, 클러스터 분석

학번: 2013-21869

목 차

표 목차

그림 목차

Ⅰ. 서 론

1. 연구의 배경 및 필요성

현재 지방자치단체의 보건사업계획 수립이 의무화(지역보건법, 1995)되면서 지역 단위 보건통계 생산의 중요성이 커지고 수요가 증가하고 있지만 지역보건정책 수립을 위한 지역단위의 대표 통계가 없었으며 '국민건강영양조사'와 같은 국가단위의 보건통계로는 한계가 있었다(김종희, 2008).

질병관리본부는 이러한 한계를 해결하기 위해서 지역 간 비교가 가능하고 지역주민의 건강수준과 보건 의식 행태 등을 파악 할 수 있도록 2007년에 20개 시·군·구를 대상으로 시범사업을 하여 2008년부터 전국 시·군·구로 확대하여 매년 지역사회건강조사를 실시하고 있다(김종희, 2008).

지역사회건강조사는 시·군·구 단위의 주민 건강통계를 생산하기 위한 우리나라 최초의 조사 사업으로 질병관리본부와 광역자치단체 및 전국 253개 시·군·구 보건소가 역할을 분담하여 지역단위 건강조사를 효율적으로 수행해오고 있다. 각 시·도의 253개 보건소에서는 지역 내 책임대학교와 협력하여 조사 수행 팀을 구성한 후 조사를 수행한다. 전국 35개 책임대학교가 평균 7개(4-14개) 보건소 관할 지역의 조사를 수행하였고 지역사회건강조사의 원활한 수행과 질 보장, 결과의 활용 제고, 지역사회의 자체 역량 강화를 달성하기 위해 노력하고 있다.

지역사회건강조사는 건강 설문조사 방법으로 수행되고 다양한 종류의 수행기관과 인력이 동시에 투입되기 때문에 자료 생산의 정확성과 신뢰성을 위해서는 조사과정 전반에 걸친 질 관리(Quality Control)가 무엇보다도 중요하다.

따라서 본 연구는 지역사회건강조사에서 발생 가능한 오류를 최소화하고 자료에 대한 객관적 검증체계를 구축하여 조사 자료의 정확성과 신뢰성을 확보하는 것을 목

적으로 한다. 구체적으로 지역사회건강조사를 수행하는 주체인 책임 대학교, 보건소 및 조사원의 조사수행 과정을 모니터링하고 질 관리(Quality Control) 정도를 평가함으로써 지역사회건강조사의 시작과 과정, 결과 등 모든 과정의 정확성과 신뢰성을 확보하고자 함이다.

2. 연구목적

전통적 의미에서 질(Quality) 좋은 통계란 "정확하고 신속한 통계"라고 강조되어 왔다. 하지만 현대적 의미에서 질(Quality)의 개념에 점차 고객만족(Customer Satisfaction)의 개념이 도입되기 시작하면서 통계자료 질(Quality)의 의미도 "통계가 얼마나 이용자들에게 사용 적합하도록 작성 및 제공되고 있는가?" 라는 과정이 부각되기 시작하고 있다. 이러한 관점에서 통계 질(Quality)의 정의는 "통계 자료가 얼마나 이용자에게 이용하기 적합(Fitness for users)하게 작성 및 제공되고 있는가를 나타내는 특성"으로 나타낼 수 있다 (Statistics Canada, 1998; 통계청, 2001).

앞에서 정의한 통계자료의 질(Quality) 관리가 이루어지지 못하여 이용자에게 부적합한 통계가 생산된다면 이용자들에게 혼란을 야기 할 뿐만 아니라 예산 낭비를 초래하게 되므로 (통계청, 2006) 통계자료의 질(Quality) 관리를 소홀히 해서는 안 된다. 또한, 정확하고 신속한 통계 결과에 국한되었던 과거와는 달리 최근에는 다양한 통계가 생산되고 이용자 요구 또한 다양화됨으로써 통계 결과뿐만 아니라 생산과정에서의 타당성과 신뢰성을 확보하기 위하여 통계 자료의 질(Quality) 관리가 요구된다.

조사통계 자료의 질(Quality)을 관리하기 위해서는 자료 내의 모든 가능한 오차를 검토하고 표본추출, 조사표설계, 자료수집 등 통계작성 절차 별로 오차를 감소시키기 위한 일련의 대책이 필요하다(Korean Statistical Society, 조사원의 업무할당 및

인구통계학적 특성에 따른 오차분석, 2004). 계속 반복하여 실시되는 통계 조사에서 조사원은 통계의 질(Quality) 관리에서 핵심적인 요인이지만 그 동안 학문적으로나 정책적으로 우리나라에서 관심의 대상이 되지 못하였다. 통계 데이터 질(Quality)은 응답자의 애매하고 사실이 아닌 답변을 하거나 설문 문항 자체의 결함에 의해 영향을 받을 수 있을 뿐 아니라 인터뷰 과정에서 조사원에 의해서도 영향을 받을 수 있다. 만약 조사원이 의도적으로 본래의 조사 규정과는 다르게 진행하는 경우 위조행위 조사(Interviewer falsification) 또는 부정행위 조사(Interviewer cheating)로 정의된다(Schreiner et al., 1988 ; Schrapler and Wagner, 2003).

부정행위는 많은 방법들에 의해 발생할 수 있는데, 정해진 가구 대상자를 방문하지 않는 경우 또는 반드시 대면조사(face-to-face interview)를 해야 함에도 불구하고 전화를 이용해서 조사를 실시하는 경우 등이 있다. 가장 심각한 부정행위는 각각의 가구 대상자를 한 번도 만나지 않고 전체 조사 내용을 위조하는 행위이다. 이러한 위조된 내용으로 발생하는 설문조사 데이터의 통계 결과는 심각한 문제로 될 수 있다.

Schnell(1991)과 Schrapler & Wanger(2003)의 의하면 위조된 조사가 전체 조사에 작은 부분으로 이루어져도 다변량(multivariate) 분석에서 큰 편향(bias)를 야기시킬 수 있다고 한다. Schrapler & Wanger(2003)에 의하면 전체 조사의 2.5% 보다 작은 위조 데이터를 포함하여 분석한 다변량 회귀(multivariate regression)의 결과가 약 80% 정도의 효과(effect) 감소가 발생했다고 한다. 이 결과를 통해서 전체 자료에서 차지하는 위조한 데이터의 심각성을 알 수 있다(Bredl et al., 2004).

부정행위 조사를 판별하는데 가장 흔한 방법은 재조사(re-interview)이다(Biemer and Stokes, 1989). 재조사 과정에 감독자를 두어 조사 참여 대상자를 직접 만나 해당 조사원에 의해 실제로 조사에 참여 했는지를 확인하며 관리하게 한다. 그러나 이런 과정으로 설문 조사에 참여한 모든 가구의 대상자들을 재조사하는 것은 비용적・시

간적으로 불가능하다. 그러므로 부정행위를 할 가능성이 있는 조사원을 탐색할 수 있는 최적의 샘플링을 해야 한다. 일반적으로 재조사할 가구를 선택하기 위해서는 부정 행위를 한 조사원 개인의 특성 또는 부정조사를 한 조사원에 의해 작성된 설문 응답의 특성이 유사한 가구를 샘플링 하는 것이 유용하다. 이때 부정 행위를 한 조사원을 '의심 (at risk)' 조사원으로 정의한다(Hood and Bushery, 1997).

조사원은 조사문항에 적혀 있는 내용을 단순히 읽어주기만 하는 것이 아니라 응답자로 하여금 조사에 성실히 임하도록 동기를 부여해 주고, 충분한 응답을 얻을 수 있도록 자세히 물어보기를 해야 한다. 응답하기 곤란한 질문에 대해서는 솔직한 답변을 이끌어낼 수 있도록 응답자들을 잘 설득해야 하지만, 이러한 과정이 지나쳐서 조사원의 편견이 개입되어서도 안 된다. 이와 같이 조사원은 설문조사에 결정적인 역할을 하기 때문에 이들이 어떤 특성을 갖고 있으며 응답태도에 어떤 영향을 미치는지를 분석하는 것은 매우 중요하다(신선옥, 2008).

통계적 접근을 통해 부정조사를 탐색한 이전의 연구가 있다(Hood and Bushery, 1997; Diekmann, 2002; Schraplera and Wagner, 2003; Swanson et al., 2003; Schafer et al., 2005). 본 연구 목적은 이전 방법으로 조사원에 의해 생성된 설문조사 데이터로 통계적 접근을 통해 부정 조사를 한 조사원을 탐색하고 설문조사에 미치는 영향을 분석하기 위함이다. 하지만 이전 연구 방법에서 조사원 탐색에 한 가지 지표(indicator)를 사용하였지만 본 연구는 이전의 연구보다 체계적인 분류를 위해서 여러 개의 지표를 합쳐서 클러스터링 분석(Cluster analyses)을 하였다.

Ⅱ. 연구 방법

1. 연구 대상

본 연구는 질병관리본부 주관으로 시행된 2012년 지역사회건강조사 원시자료(raw data), 2012년 지역사회건강조사 폐기자료, 2012년 지역사회건강조사 질(Quality) 관리 지표와 조사원 특성 자료를 이용하였다. 2008년부터 전 국민을 대상으로 시행하는 지역사회건강조사는 질(Quality) 향상을 위해 전화 점검을 실시하여 조사의 신뢰도를 향상시키고 조사의 품질을 관리하기 위해 노력하고 있다. 전화 점검은 제 3의 기관에서 이루어지고 전체 조사완료 대상자 약 22만명의 10%인 22,867 샘플을 무작위 추출하여 2012년 8월 중순부터 11월 말까지 진행하였으며 57,273건의 전화 시도로 총 25,138 샘플의 전화점검을 완료하였다. 2012년 전화 점검 문항은 총 11개로 조사과정의 진실성 검증을 위해 실제 조사 참여 여부 문항(1개)과 CAPI(Computer-Assisted Personal Interviewing) 사용 여부 문항(1개)이 포함되어 있다. 또한, 조사문항의 임의 작성 검증을 위해서 지역사회건강조사의 결과와 전화점검 결과의 일치도 평가 문항(6개)이 포함되어 있으며 나머지는 조사원 친절도 문항(1개), 답례품 정상 수령 여부 문항(1개), 성별에 대한 문항(1개)이 포함되어 있다. 전화 점검의 결과를 통해 의심되는 건에 대해서는 책임대학교에 알리고 진위 여부를 확인하며 조치사항에 대한 피드백이 이루어 진다. 이러한 전화점검 결과는 질(Quality) 관리 평가에도 반영하여 조사가 원활히 이루어 질 수 있게 하였다.

질(Quality) 관리 평가 지표 중 본 연구에서는 조사원 별 표본가구대체율, 가구완료율, 응답거부율, 응답일치도, 조사 소요시간 이상치 비율의 데이터를 이용하였고 <Table1> 각 지표의 결과는 <Table2>와 같다. 총 1413명의 조사원 특성

자료는 조사원의 연령, 연차, 업무시작 시기, 조사완료 여부, 부정행위 여부, 조사자료 폐기 여부가 포함되어 있다 <Table3>. 업무시작 시기는 조사원이 조사를 하기 전 의무적으로 교육을 받게 되는데 조사원 교육 이전에 시작 하는 경우, 교육은 받지 못했지만 조사를 시작 하기 전에 업무를 시작하는 경우, 조사 중에 업무를 시작하는 경우로 나눌 수 있다. 조사 중단 여부는 업무를 시작한 이후에 조사를 완료한 경우, 조사를 포기하는 경우, 해고를 당하는 경우로 나눌 수 있다. 부정행위의 구체적인 예로 허위로 조사를 하거나 대리 응답을 하여 적발 되는 경우 등이 있고 자료를 폐기하는 것은 부정행위가 적발 되거나 조사 대상자가 아닌데 조사를 하는 경우 등이 있다. 총 1413명의 조사원이 2012년 지역사회건강조사에 참여하였고 전체 조사원의 평균 연령은 45.1세, 조사원 평균 연차는 2.1년이었다. 지역사회건강조사를 수행한 시기는 조사원 교육이전이 77.9%로 가장 많았고 조사시작 전에 수행한 조사원이 11.9%, 조사 중에 업무를 시작한 조사원이 10.2%였다. 지역사회건강조사의 중단 여부는 전체 조사원 중 조사를 완료한 조사원이 93.7%, 조사를 포기한 조사원이 77명으로 5.5%를 차지하였고 해고를 당한 조사원이 12명으로 전체 조사원의 0.9%였다. 또한, 부정행위를 한 조사원이 50명으로 전체 조사원 중 3.5%를 차지했고 자료폐기를 한 번이라도 한 경험이 있는 조사원이 140명으로 전체 조사원 중 9.9%를 차지한다.

Table 1 질 관리 평가 지표

지표명	지표정의	산출식
표본가구대체율	표본가구 중 아래 사유로 대체된 가구의 분율 [표본가구대체율 포함: 조사 적합 사유] ① 3회 이상 방문: 접촉불가(성인거주) ② 3회 이상 방문: 조사거부 ③ 기타 부적합사유 ④ 오입력자료	$\frac{\text{표본가구대체율}}{\text{표본가구수}} \times 100$
가구완료율	조사가구 내 만19세 이상 성인 중 조사적격 가구원을 모두 정상 완료한 가구의 분율	$\frac{\text{조사완료가구수}}{\text{조사대상가구수}} \times 100$
응답거부율	응답자 중 가구소득, 키, 몸무게, 학력에 대한 응답거부 분율의 평균	$\frac{\text{문항별 응답거부분율의 합}}{4} \times 100$
응답일치도	본조사 응답과 전화확인 조사 결과의 일치도(6문항) **<일치도 확인문항>** 문항1. 평생 음주 여부 문항2. 지난 1년간 음주 여부 문항3. 자동차 운전 여부 문항4. 최근 1년간 스케일링 여부 문항5. 최근 1년간 독감예방접종 여부 문항6. 당뇨병 의사진단 여부	전화점검조사 6문항 응답값과 본조사자료 응답값이 동일한 항목의 평균 개수
조사소요시간 이상치 비율 (9분이내)	조사소요시간이 9분 이내로 다른 조사자료의 조사소요시간보다 월등히 짧은 경우	$\frac{\text{9분미만 개인조사자료수}}{\text{120분미만 개인조사자료수}} \times 100$

Table 2 질 관리 평가 지표 요약

질 관리 지표	N (명)	Mean	SD	Min	Q1	Q2	Q3	Max	Range
표본가구대체율 (%)	8,651	3.96	2.97	0	1.6	3.4	5.6	14.6	14.6
가구완료율 (%)	220,165	92.6	6.06	69.1	89.2	94.1	97.1	100	30.9
응답거부율 (%)	5,303	0.99	1.13	0	0.18	0.56	1.43	5.66	5.66
응답일치도 (개)	25,138	5.49	0.18	4.90	5.39	5.53	5.62	5.82	0.93
조사소요시간 이상치 비율 (%)	221,052	0.64	1.54	0	0	0.11	0.57	12.9	12.89

Table 3 조사원 특성

변수	분류	N (명)	비율 (%)
연령	missing	45	3.2
	20~29	10	0.7
	30~39	187	13.2
	40~49	879	62.2
	50~59	285	20.2
	60~69	7	0.5
연차	missing	32	2.3
	1년	699	49.5
	2년	280	19.8
	3년	134	9.5
	4년	140	9.9
	5년	128	9.1
업무시작시기	조사원교육 이전	1101	77.9

	조사 시작 전	168	11.9
	조사 중	144	10.2
조사중단여부	조사완료	1324	93.7
	조사포기	77	5.5
	해고	12	0.9
부정행위 여부	없음	1363	96.5
	있음	50	3.5
자료 폐기 여부	없음	1273	90.1
	있음	140	9.9

2. 통계적 분석 방법

2.1 Benford’ s Law

Benford 법칙이 처음으로 알려진 것은 1881년 미국의 수학자인 Simon Newcomb(1881)에 의해서이며 로그표(logarithm table)의 책들이 항상 앞 페이지만 지저분하고 뒤로 갈수록 깨끗하다는 사실을 발견하였다. 로그표는 수가 점점 커지는 순서대로 배열되어 있기 때문에 맨 앞자리수가 작은 숫자가 큰 숫자에 비해 더 많이 쓰인다는 사실을 발견하였다. 즉, 첫 번째 유효숫자(first significant digit or leading digit) 1 또는 2로 시작하는 숫자는 8 또는 9 보다는 더 자주 본다는 것이다. 그 후 1938년에 Frank Benford 라는 물리학자가 강의 너비, 사망률, 잡지에 있는 숫자, 야구 통계 등 지형학적・인구학적인 다양한 데이터로부터 임의의 20,229개 숫자를 분석하여 맨 앞자리 수가 ‘1’일 확률이 1/9이 아니라 0.31, ‘2’가 0.19, ‘9’가 0.05의 확률이라는 Benford 법칙을 논리적으로 재발견하였고 첫 번째 유효숫자(first digit)에 따른 확률은 <Table 4>와 같다

Table 4 Benford 법칙의 확률분포표

First digit	1	2	3	4	5	6	7	8	9
p	0.30103	0.17609	0.12494	0.09691	0.07918	0.06695	0.05799	0.05115	0.04578

Hill(1995)은 “어떤 분포를 임의로 골라서 이 분포들에서 임의로 자료를 모으면 각

분포들 자체는 아니더라도 결합된 자료의 분포는 Benford 법칙을 따른다" 는 것을 이론적으로 증명하였고 사회·경제적 수치와 재무회계 분야에도 적용하기 시작하였다. Varian(1972)에 따르면 이 이론이 절대적인 법칙은 아니지만 회계 수치의 부정에 대한 1차적 검증의 기능이 있으므로 자료의 진실성을 알 수 있다고 한다. Benford 법칙은 위·변조 방지(detecting fraud)와 신뢰성 검증(checking reliability)을 위한 간접적 도구(Gunnel & Todter, 2009)로 사용될 수 있으며, 모형 결과를 조사하는 진단도구(Varian, 1972)와 부정행위의 지표(Todter, 2009)로 사용될 수 있다.

Schrapler and Wagner(2003) 와 Schafer et al(2005)의 논문에서 독일 사회·경제적 패널 조사(SOEP)의 부정행위를 한 조사원을 탐색하기 위해 Benford 법칙이 사용되었다. 두 연구에서 조사원 별로 모든 설문지를 합치고 조사원 별로 첫 번째 유효숫자(first digit)가 Benford의 법칙을 따르며 유의하게 분포하고 있는지를 확인하였다. 이것은 x^2 통계량으로 이루어 졌으며 식은 (1)과 같다.

$$x^2{}_i = n_i \ \sum_{d=1}^{9} \frac{\left(h_{d_i} - h_{b_d}\right)^2}{h_{b_d}} \qquad (1)$$

식 (1)에서 n_i 는 조사원 i 별 모든 설문지의 첫 번째 유효숫자(first digit)의 개수(number)이다. h_{d_i} 는 Benford 분포에서 조사원 i 별 첫 번째 유효숫자(first digit) d=1,...,9 의 비율을 나타내고 h_{bd} 는 Benford's law 분포에 따른 d=1,...,9 의 비율이다. x^2값이 높은 것은 Benford 분포와의 편차(deviation)가 있다는 것을 의미하고 이것은 의심 조사원(at risk interviewer)으로 분류된다.

Benford 법칙은 특정한 조건에서만 유효하며 이 법칙을 이용하여 설문조사의 위조된 데이터(falsified data)를 추출하기 위해서는 몇 가지 충족해야 하는 조건들이 존재한다. 이러한 필요 조건들은 모의실험 결과(Scott/Fasli 2001)이며, 나머지는 실제 응용(Nigrini 1999) 또는 이론적 분석(Hill 1995)으로 도출하였다.

- 데이터는 정해진 최대값(built-in maximum)이 존재하지 않아야 한다. 왜냐하면 최대값이 존재하면 자릿수에서 더 많이 발생하는 빈도의 숫자가 발생할 것이고 이것은 편향된 결과를 유발한다.
- 데이터는 개인 신원 번호(social security number)나 은행 계좌와 같이 할당된 숫자를 포함할 수 없다(Nigrini, 1999).
- 데이터는 0이 아닌 단봉분포(unimodal distribution)의 양의 값을 가져야 한다 (Scott/Fasli, 2001).
- 데이터는 다른 데이터로부터 계산된 평균 또는 분산과 같이 통계 과정으로부터 도출된 것이 아니어야 한다(Mochty, 2002).
- 설문 조사에서의 범주형 자료(Categorical data)는 위에서 언급한 조건을 충족하지 않으며 오직 연속형 변수(continuous variables)만 가능하다.

또한, 데이터 표본의 크기가 충분히 클수록 위의 필요 조건을 더 잘 만족하며 Benford 분포를 충족한다(fit of Benford's distribution).

본 연구에서는 Schrapler and Wagner(2003) 와 Schafer et al(2005)의 연구와 동일하게 Benford 분포를 이용해 정상적으로 조사한 조사원(honest interviewer)과 부정행위를 한 조사원(cheating interviewer)들의 첫 번째 유효숫자(first digit) 분포의 x^2값을 각각 구하였다. Beford 법칙 분석을 위한 지역사회건강조사의 문항은 다음과 같다

- 가구소득(금액): 임금, 부동산 소득, 연금, 이자, 정부 보조금, 친척이나 자녀들의 용돈 등 모든 수입을 합쳐 최근 1년 동안 가구의 총 소득
- 하루 평균 흡연량(개비): 하루 평균 흡연량, 흡연하는 날 하루 평균 흡연량,

과거 담배를 피울 때 하루 평균 흡연량

<Figure1>과 <Figure2>는 지역사회건강조사 본 조사 데이터와 부정행위를 하여 폐기된 조사 데이터의 첫 번째 유효숫자(first digit)의 분포를 나타낸 것으로 꺾은 선 그래프는 Benford 법칙의 분포를 나타내었다. 두 데이터 모두 Benofrd 법칙 분포와 다른 점은 첫 번째 유효 숫자에서 '2'의 비율이 높은 점이다. 폐기된 조사 데이터에서 <Figure4>의 가구 소득의 경우 숫자 '3'의 비율이 높고 첫 번째 유효 숫자가 커질 수록 비율이 줄어드는 Benford 법칙과는 다른 결과를 나타낸다. 또한 <Figure6>에서 하루 평균 흡연량의 첫 번째 유효 숫자 '8'의 비율이 높아지며 Benford 법칙 분율을 나타낸 꺾은 선 그래프와 차이가 나는 점을 확인 할 수 있다.

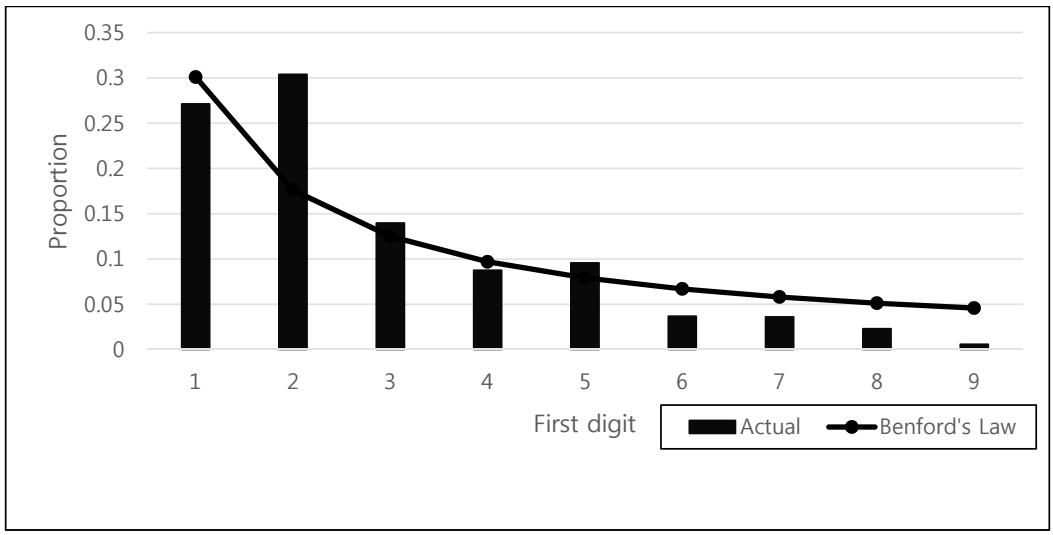

Figure 1 First-digit distribution of all numbers in the cases of true data (n=159994)

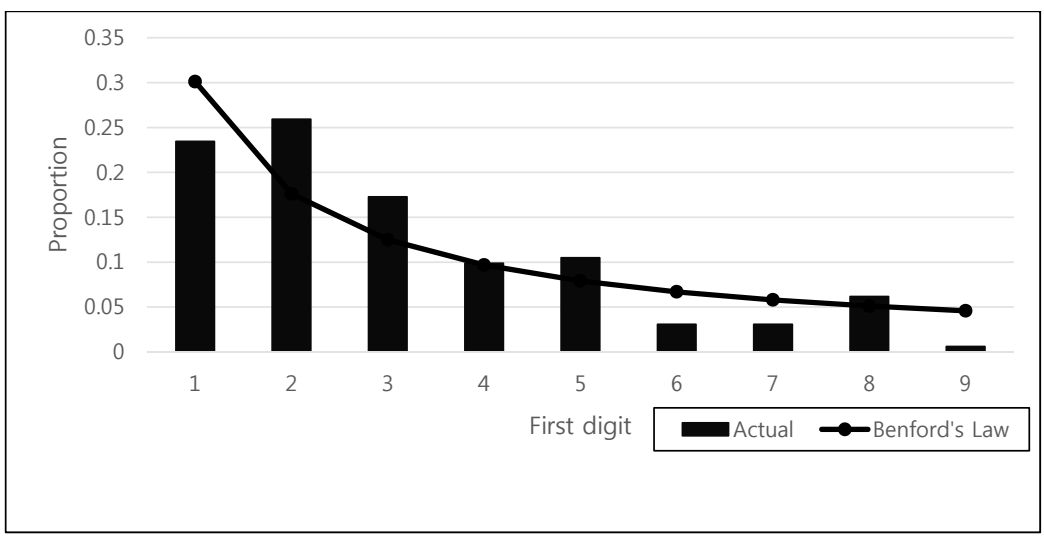

Figure 2 First-digit distribution of all numbers in the cases of faked data (n=162)

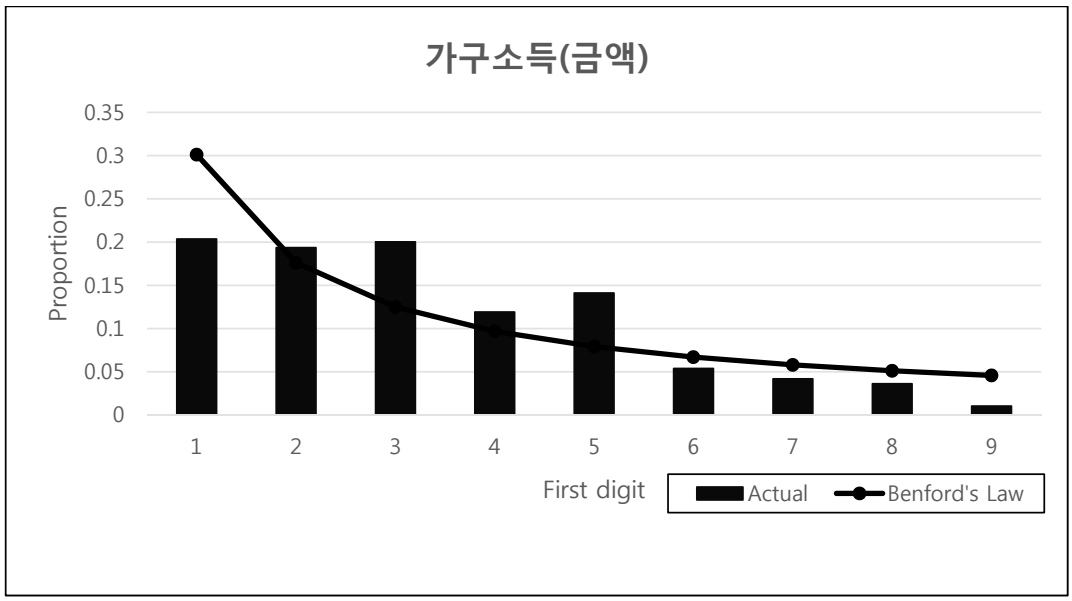

Figure 3 First-digit distribution of all numbers for the item "family income" in the cases of true data (n=79947)

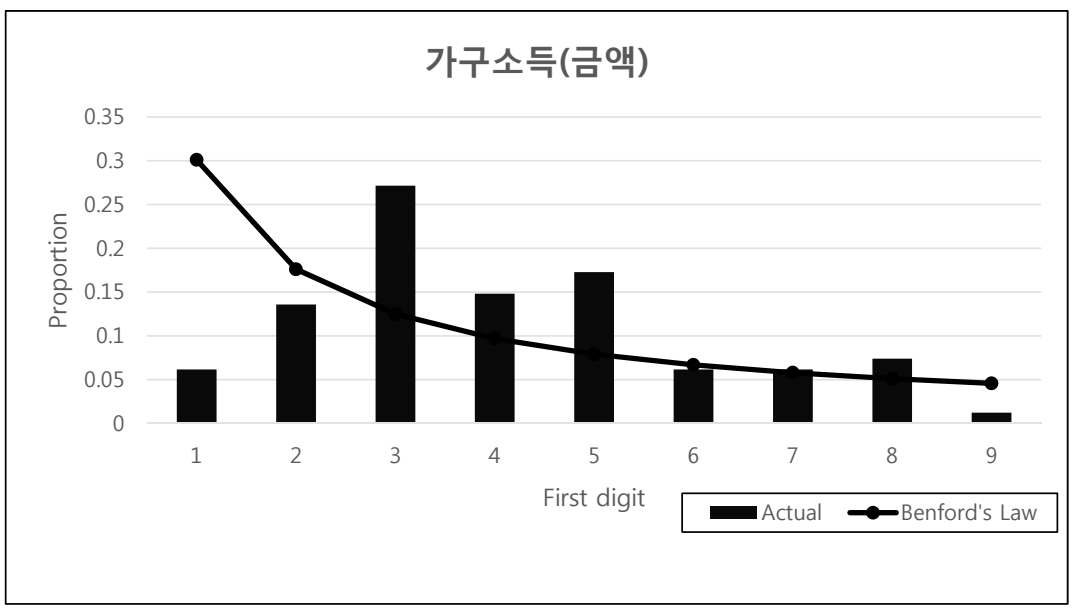

Figure 4 First-digit distribution of all numbers for the item "family income" in the cases of faked data (n=81)

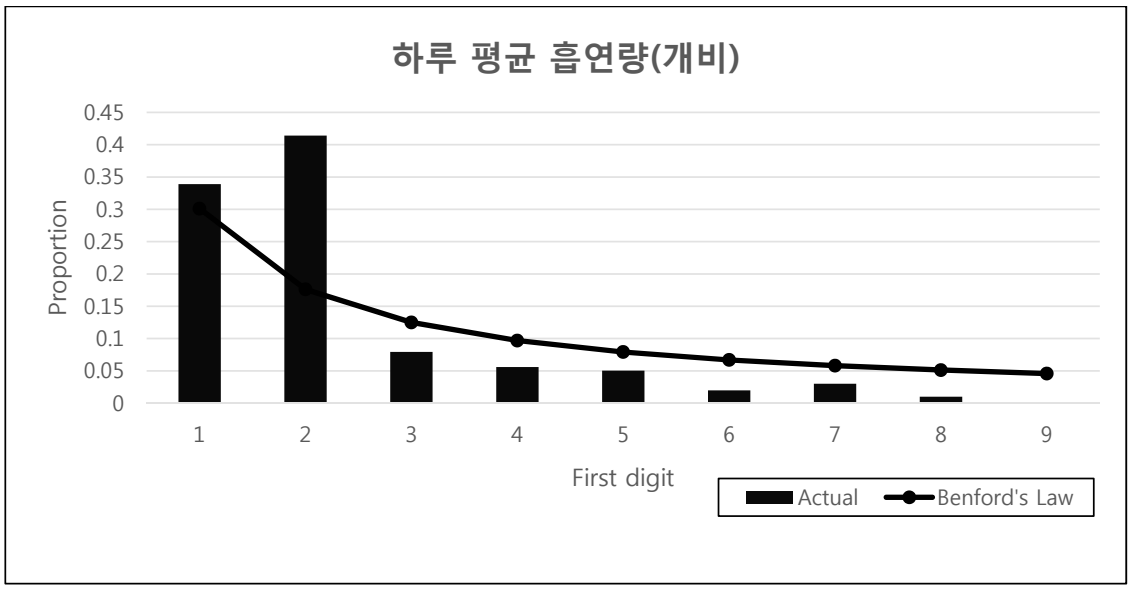

Figure 5 First-digit distribution of all numbers for the item "Number of cigarettes smoked in a day" in the cases of true data (n=80047)

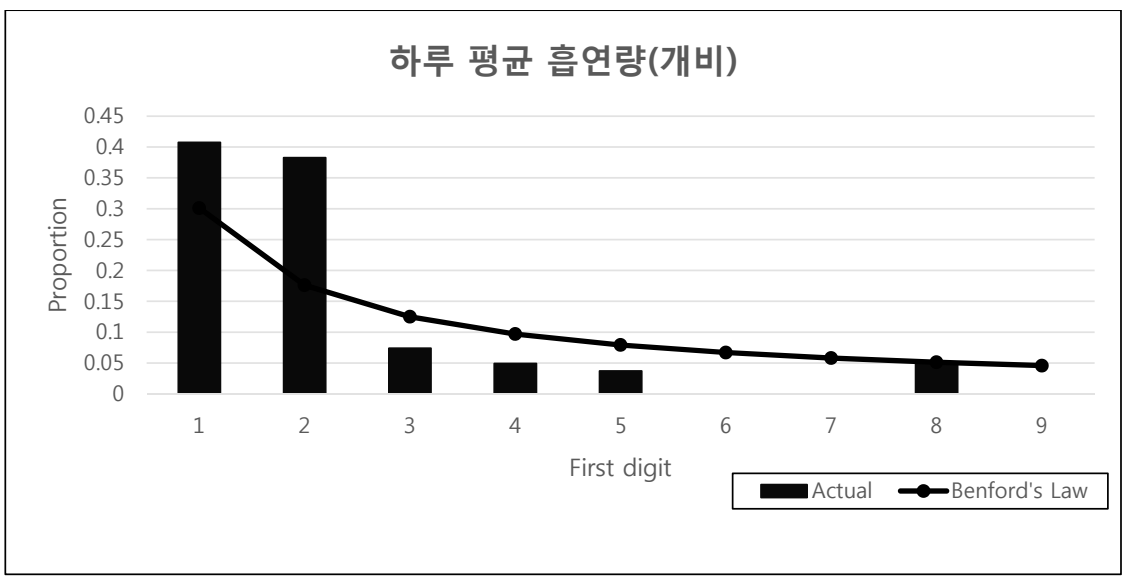

Figure 6 First-digit distribution of all numbers for the item "Number of cigarettes smoked in a day" in the cases of faked data (n=81)

2.2 다변량 분석법(Multivariate Analysis)

각 조사원들의 설문지로부터 부정행위 여부를 알 수 있는 지표(indicators)들을 이용해 정상 조사원의 클러스터 분석(cluster analysis)을 하였다. 이를 통해 지표들에 의해 분류된 조사원과 기존에 분류된 정상 조사원(honest interviewer)과 부정행위를 한 조사원(cheating interviewers)을 비교 할 수 있었다.

클러스터 분석은 다섯 개의 지표를 이용하여 두 개의 그룹으로 나뉘는데 한 그룹은 부정행위를 할 가능성이 있는 조사원(at risk interviewers)이 포함된 그룹으로 분류되고 다른 한 그룹은 정상 조사원(honest interviewers)이라고 가정하는 그룹으로 분류된다. 사실 이러한 방법은 조사원 중 누가 부정행위를 했는지 사전의 정보를 가지는 것이 필수는 아니며 클러스터 방법을 통해 알게 된다는 것을 가정한다. 하지만 부정 행위를 한 조사원이 누구인지 이미 알고 있다면 클러스터 방법이 “실제로” 정상적으로 조사를 한 조사원(honest interviewer)이 부정행위를 한 조사원(cheating interviewers)으로 분류 한 것을 확인할 수 있고 반대로 부정행위를 한 조사원(cheating interviewers)가 정상적으로 조사를 한 조사원(honest interviewer)로 분류된 것을 확인 가능하다.

첫 번째 지표는 앞에서 언급한 Benford 분포와 각 조사원 설문지의 첫 번째 유효숫자(first digit)의 분포를 비교한 x^2 값을 사용한다. 나머지 세 개의 지표는 위조된 데이터(fabricating data)의 부정 행위를 한 조사원들의 행동에 관한 가설을 언급한 논문(Schafer et al., 2005)에 따르면 부정 행위를 하는 조사원(cheaters)들은 모든 질문에 대답하는 경향이 있기 때문에 결측값이 적을 것이라고 가정하였다. 또한 그들은(cheaters) 순서형(ordinal) 문항에서 극단적 답변(extreme answers)을 적게 할 것이라고 가정하였다. Hood and Bushery(1997)에 따르면 “부정 행위를 하는 조사원

(cheaters)들은 문항을 되도록이면 적게 답하려고 노력할 것이고, 위조된 데이터를 최소화 하면서 위조할 것이다" 라고 가정하였다. 이러한 가정을 바탕으로 네 개 지표들의 조사원 별(interviewer-level) 비율을 구하였다.

- 응답거부율(non-response-ratio): 응답자 중 '가구소득, 키, 몸무게, 학력'에 대한 응답거부 비율의 평균이다. 부정 행위를 하는 조사원(cheaters)은 정상 조사원(honest interviewer)에 비해 낮을 것으로 가정한다.

- 극단적 응답 비율(extreme-answers-ratio): 순서형(ordinal) 문항의 답변에서 양 극단에 있는 답변을 응답한 비율이다. 예를 들어, 지역사회건강조사 문항 중 '주관적 구강건강 수준'에서의 응답 '매우 좋음, 좋음, 보통, 나쁨, 매우 나쁨' 다섯 가지 중 '매우 좋음'과 '매우 나쁨'으로 응답한 비율을 의미한다. 극단적 응답 비율은 '음주빈도, 자동차 안전벨트 착용, 영양표시 확인, 주관적 구강건강 수준, 저작불편 호소, 발음불편 호소'에 대한 극단적 응답 비율의 값이다. 부정 행위를 하는 조사원(cheaters)은 정상 조사원(honest interviewer)에 비해 극단적 응답률이 낮을 것으로 가정한다.

- 기타 응답 비율(other-answers-ratio): 설문지의 객관식 답변 중 대상자가 원하는 답변이 없는 경우 기타 표시 후 주관식으로 직접 기입한 비율이다. 기타 응답 비율은 '주거형태, 금연이유, 필요 치과진료 미수진 사유, 건강검진 미수진 사유, 암검진 미수진 사유, 필요의료서비스(치과제외) 미수진 사유, 보건기관 이용 이유, 보건기관을 이용하지 않는 이유'에 주관식으로 응답한 비율의 값이다. 기타 응답 비율은 조사의 수고로움을 덜기 위하여 '부정 행위를 하는 조사원(cheaters)은 정상 조사원(honest interviewer)에 비해 기타 응답률이 낮을 것으로 가정한다.

- 여과문항 응답 비율(filter-question-answers-ratio): 여과문항은 그 응답에

따라 다음 문항을 응답 여부가 결정되는 것으로 예를 들어, 여과 문항에서 ‘아니오’라고 답변을 하면 비해당자가 되어 다음 문항을 생략 가능하다. 이때, 여과문항 응답 비율은 여과문항을 이용하여 다음문항을 생략한 비율을 의미한고 지역사회건강조사의 총 40개 문항의 여과 문항 응답 비율을 이용하였다. 부정 행위를 하는 조사원(cheaters)은 시간을 절약하기 위해 여과문항에서 ‘아니오’를 이용하여 다음 문항을 생략할 비율이 높다고 가정한다.

Ⅲ. 연구 결과

1. 조사원 분류

2012년 지역사회 건강조사의 전체 조사원 중 결측값을 제외한 총 1219명 조사원의 질관리 지표와 조사원 특성 자료를 얻을 수 있었다. 같은 방법으로 연구한 선행 논문에서는 의심 조사원(at risk interviewer)을 탐색하기 위해 재조사(re-interview)를 통해서 부정행위를 적발하고 적발된 조사원의 설문 응답을 바탕으로 분석을 하였다. 이 때의 재조사는 감독자(supervisor)를 동반하여 조사하고자 하는 대상자를 방문하여 똑 같은 설문을 한 번 더 실시하는 것이다. 하지만 지역사회건강조사의 재조사 방법은 기존 연구와는 다르게 전화조사로 재조사가 실시되며 전체 조사원 중 임의적으로 10%를 추출하여 전체 문항이 아닌 6개의 문항의 일치도 분석을 실시한다.

전화점검을 통한 재조사에서 일치도 문항 6개가 모두 일치하고 조사원 특성 자료를 통해 부정조사를 한 경험이 한 번도 없으며 자료 폐기의 경험이 한 번도 없는 경우를 정상 조사원(honest interviewer)이라고 분류하였다. 또한 폐기된 자료를 통하여 대리응답이나 조사원을 직접 만나지 않고 조사를 수행한 경우 부정행위를 한 조사원(cheating interviewer)로 분류하였다. 폐기된 자료는 전화점검을 통한 재조사로 '지역사회건강조사 미참여 확인', 'CAPI 미사용', '답례품 미제공'의 조사원 자료이며 자체 대학에서 재확인을 통해 부정 행위 여부를 확인한다.

폐기 자료에서 부정행위 중 대리조사가 확실하게 명시된 10명의 조사원(cheating interviewer)과 전화조사와 본 조사의 응답일치도가 6이며 폐기자료 건수가 0이고 부정행위를 단 한번도 경험하지 않은 12명의 조사원(honest interviewer)을 대상으로

분석하였다.

<Table5>는 정상 조사원(honest interviewer) 12명과 부정행위를 한 조사원(cheating interviewer) 10명의 조사를 완전하게 완료한 건수를 나타낸다. 부정행위를 한 조사원(cheating interviewer)의 조사완료 건수는 총 36건(총 10명), 정상적으로 조사를 한 조사원(honest interviewer)의 조사완료 건수는 총 510건(총 12명) 이다. 지역사회건강조사는 조사를 진행하는 도중에 전화 점검을 통해 재조사를 수행하기 때문에 대리응답과 같은 부정조사를 하는 경우 조사 중에 적발되어 자료는 폐기되고 조사원은 조사 후 대리응답이 확인되면 경고 및 주의 조치를 받거나 해고되는 경우가 대부분이다. 그러므로 정상적으로 조사를 수행한 조사원(honest interviewer)과 부정행위를 한 조사원(cheating interviewer)의 조사 완료 건수는 차이가 난다.

대리응답으로 부정행위를 한 조사원(cheating interviewer) 중 C5가 7건으로 조사 완료 건수가 가장 많고 C7이 1건으로 가장 적은 반면에 정상적으로 조사를 수행한 조사원(honest interviewer)의 조사완료건수는 H10이 73건으로 가장 많으며 H6이 12건으로 가장 적게 조사를 하였다.

Table 5 부정행위를 한 조사원(cheating interviewer)과 정상 조사원(honest interviewer)의 조사 완료 건수

Cheating interviewer	C1	C2	C3	C4	C5
N	2	5	6	3	7
Cheating interviewer	C6	C7	C8	C9	C10
N	3	1	2	5	2
Honest interviewer	H1	H2	H3	H4	H5
N	19	65	40	45	44
Honest interviewer	H6	H7	H8	H9	H10
N	12	54	35	19	73
Honest interviewer	H11	H12			
N	44	60			

2. 클러스터 분석(cluster analysis)

조사원 분류 단계에서 질 관리 지표의 응답일치도와 조사원 특성의 부정행위 여부, 자료폐기 여부를 이용해서 정상 조사원(honest interviewer)과 부정행위를 한 조사원(cheating interviewer) 두 그룹으로 클러스터링(clustering)을 하였다. <Table6>에서 부정행위를 한 조사원(cheating interviewer)의 지표값은 기타 응답 비율(other-answers-ratio)이 모두 0이며 순서형(ordianal) 문항에서 극단적 응답을 한 비율(extreme- answers-ratio)이 C2와 C3가 0이며 C10을 제외하고 50 미만의 값으로 낮다. 응답거부율(non-response-ratio)은 C8을 제외한 나머지는 1미만의 값으로 응답거부율이 낮으며 여과문항에서 '아니오' 라고 답하여 다음 문항에 응답하지 않고 생략한 비율(filter-question-answers-ratio)이 C4가 68.53으로 가장 낮은 값을 가지고 나머지 부정행위를 한 조사원은 80이상의 값을 가진다. Benoford 카이제곱의 값은 C9를 제외하고 100이상의 값을 가진다. <Table 6>

폐기된 자료가 없고 부정행위를 하지 않은 조사원 중에서 응답일치도가 6으로 재조사에서 일치 문항을 모두 만족한 조사원(honest interviewer)의 기타응답 비율(other-answers-ratio) 은 0인 조사원이 4명이며 나머지는 H5는 18.18의 값을 가진다. 순서형(ordinal) 문항에서 극단적 응답을 한 비율(extreme- answers-ratio)은 H11가 63.64로 가장 컸으며 응답거부율(non-response-ratio)은 H4, H9, H10을 제외하고 나머지 모두 1 미만이었다. Benford 카이제곱값은 H2, H3, H7 H12를 제외한 나머지는 100 미만의 값을 나타냈고, 여과문항에서 '아니오' 라고 답하여 다음 문항에 응답하지 않고 생략한 비율(filter-question-answers-ratio)이 H1과 H5를 제외하고 80미만의 값을 가진다. <Table 7>

Table 6 부정 행위를 한 조사원(cheating interviewer)의 지표값

조사원	기타 응답 비율 (%)	극단적 응답 비율 (%)	응답거부율 (%)	Benford x^2	여과문항 응답율 (%)
C1	0	0	0.51	261.13	87.5
C2	0	0	3.8	349.51	85
C3	0	0.21	0	152.09	82.5
C4	0	4.46	0	260.39	68.53
C5	0	33.33	0.16	147.27	85
C6	0	33.33	0.21	234.99	85
C7	0	33.33	0.4	431.48	77.5
C8	0	33.33	1.47	326.61	85
C9	0	16.67	0	85.04	70
C10	0	50	0	115.83	77.5

Table 7 정상 조사원(honest interviewer)의 지표값

조사원	기타 응답 비율 (%)	극단적 응답 비율 (%)	응답거부율 (%)	Benford x^2	여과문항 응답율 (%)
H1	0	14.91	0	13.70	82.37
H2	0	15.13	0	164.44	78.65
H3	0	18.33	0	129.22	75.44
H4	0.56	25.19	1.15	47.68	78.5
H5	18.18	33.33	0	51.73	80.34
H6	17.71	36.11	0.76	63.18	78.75
H7	0.93	44.75	0	157.03	78.56
H8	1.07	46.67	0.54	27.11	75.86
H9	0	51.75	1.25	17.50	72.37
H10	0.51	58.90	1.24	59.68	73.77
H11	1.70	63.64	0.39	57.53	73.69
H12	0.21	37.22	0.51	195.25	75.42

<Table8>은 부정행위를 한 조사원(cheating interviewer) 10명과 문항 일치도의 값이 6으로 정상적으로 조사하였다고 가정한 12명 조사원(honest interviewer)의 다섯 가지 지표(기타 응답 비율, 극단적 응답율, 응답거부율, benford x^2 값, 여과문항 응답율)에 대한 평균, 최대값, 중위수, 최소값을 나타낸 것이다.

기타를 응답한 비율(other-answers-ratio)의 평균이 부정행위를 한 조사원(cheating interviewer)은 0으로 앞에서 가정한 바와 같이 정상 조사원(honest interviewer)에 비해 값이 낮았고, 극단적 응답율(extreme-answers-ratio)의 경우도 앞의 가정과 같이 부정행위를 한 조사원(cheating interviewer)의 값이 정상 조사원(honest interviewer)에 비해 낮게 나왔다.

하지만 응답거부율(non-response-ratio)은 앞의 가정과는 달리 부정행위를 한 조사원(cheating interviewer)이 정상조사원(honest interviewer) 보다 더 높게 나왔다.

benford x^2 값과 여과 문항 응답율(filter-question-answers-ratio)도 앞의 가정과 마찬가지로 부정행위를 한 조사원(cheating interviewer)이 정상조사원(honest interviewer) 보다 더 높게 나와 가정과 일치하는 결과가 나왔다.

Table 8 부정행위를 한 조사원(cheating interviewer)과 정상 조사원(honest interviewer)의 지표값

Cheating interviewer	기타 응답 비율 (%)	극단적 응답율 (%)	응답 거부율 (%)	Benford x^2	여과문항 응답율 (%)
평균	0	20.47	0.66	236.43	80.36
표준편차	0	18.42	1.1	111.89	6.71
최대값	0	50	3.8	431.48	87.5
중위수	0	25	0.19	247.69	83.75
최소값	0	0	0	85.04	68.53
Honest interviewer	**기타 응답 비율 (%)**	**극단적 응답율 (%)**	**응답 거부율 (%)**	**Benford x^2**	**여과문항 응답율 (%)**
평균	3.14	35.71	0.45	85.64	76.86
표준편차	6.59	16.72	0.51	61.18	2.91
최대값	18.18	63.64	1.25	195.25	82.37
중위수	0.51	36.11	0.39	59.68	75.86
최소값	0	14.91	0	13.7	72.37

클러스터링 분석(clustering analysis)은 다변량 기법의 하나로서 전체 데이터 집합을 유사한 성질을 갖는 몇 개의 클러스터로 분할하고 대상물들이 지니고 있는 특성을 토대로 이들을 분류한다. 이렇게 분류된 대상물들로 구성된 군집들은 내적(군집 내) 동질성이 높고 외적(군집 간) 이질성이 높아지는 결과를 보여주게 된다. 클러스터링 분석(clustering analysis)의 목적은 관측개체를 몇 개의 그룹으로 나누어 대상 집단에 대한 해석을 하고 효율적인 활용하기 위함이다.

K-평균화 클러스터링(K-means Clustering)은 비계층적 군집화의 대표적인 기법으로 N개의 속성으로 구성되는 각각의 레코드를 벡터로 표시하여 N차원의 데이터 공간에 나타낼 때, 유사한 특성을 갖는 레코드들은 서로 근접하여 위치한다는 가정에 근거하고 있다. 여기서 K는 군집수로 사전에 결정되며 K개의 평균점을 지정하고 모든 데이터에서 가장 가까운 평균점에 해당하는 그룹에 할당한다. 그 후에 다시 평균점들을 조금씩 바꾸면서 가장 가까운 그룹에 재할당하며 이 과정은 군집 변동이 없을 때까지 계속 반복된다.

```
단계 1) 처음 k개의 개체로 k개의 클러스터를 만든다.
단계 2) 남아 있는 (n-k) 개의 개체에 대하여
        가장 가까운 중심을 갖는 클러스터에 삽입한다.
        변경된 클러스터의 중심을 다시 계산한다.
단계 3) n 개의 모든 개체에 대하여
        가장 가까운 중심을 갖는 클러스터에 삽입한다.
        변경된 클러스터의 중심을 다시 계산한다.
단계 4) 모든 클러스터의 중심이 변경되지 않을 때까지 단계 3)을 반복한다.
```

Figure 7 K-means Clustering 알고리즘

<Table9>는 기타 응답 비율(other-answers-ratio)에 의한 k-means clustering의 결과이며 <Table15>에서 기타 응답 비율(other-answers-ratio)을 이용하여 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 100%로 가장 높게 나왔으며, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 83.33%라는 것을 알 수 있다.

<Table10>는 극단적 응답율(extreme-answers-ratio)에 의한 k-means clustering의 결과이며 <Table15>에서 극단적 응답율(extreme-answers-ratio)을 이용한 클러스터링 방법으로 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 50%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약 66.67%라는 것을 알 수 있다

<Table11>은 응답거부율(non-response-ratio)에 의한 k-means clustering의 결과이며 <Table15>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 약 20%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 75%이다. 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 가장 낮았으며 응답거부율을 통해 부정조사를 한 조사원(cheating interviewer)을 탐색하는 것이 어렵다는 사실을 알 수 있다.

<Table12>은 Benford x^2에 의한 k-means clustering의 결과이며 <Table15>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 약 60%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약 100%이며 Benford 분포를 이용하여 정상적으로 조사를 한 조사원(honest interviewers)를 탐색 가능 한 사실을 알 수 있다.

<Table13>은 여과 문항 응답율(filter-question-answers-ratio)에 의한 k-means clustering의 결과이며 <Table15>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 60%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 50%라는 것을 알 수 있다.

<Table15>는 기타 응답 비율(other-answers-ratio), 극단적 응답율(extreme-answers-ratio), 응답거부율(non-response-ratio), Benford x^2 , 여과 문항 응답율(filter-question-answers-ratio)의 5가지 지표를 모두 포함한 다변량

분석(multivariate analysis)을 한 결과이다. C4를 제외하고 실제로 부정행위를 한 조사원(cheating interviewers)을 탐색한 비율이 90%이며 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 100%이다.

단변량 분석(univariate analysis) 중에서 기타 응답 비율(other-answers-ratio)을 이용한 방법이 부정행위를 한 조사원(cheating interviewers)을 탐색한 비율이 높았고, Benford x^2 은 정상적으로 조사를 한 조사원(honest interviewers)을 탐색하는 비율이 가장 높았다.

Table 9 기타 응답 비율(other-answers-ratio)에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	C	C	C
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	C	C
Honest interviewer	H1	H2	H3	H4	H5
group	C	C	C	C	C
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	C	C	C
Honest interviewer	H11	H12			
group	C	C			

Table 10 극단적 응답율(extreme-answers-ratio)에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	C	C	H
Cheating interviewer	C6	C7	C8	C9	C10
group	H	H	H	C	H
Honest interviewer	H1	H2	H3	H4	H5
group	C	C	C	C	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 11 응답거부율(non-response-ratio)에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	H	C	H	H	H
Cheating interviewer	C6	C7	C8	C9	C10
group	H	C	H	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	C	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	C	C	H
Honest interviewer	H11	H12			
group	H	H			

Table 12 Benford x^2 에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	C	H
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 13 여과 문항 응답율(filter-question-answers-ratio)에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	C	H	C
Cheating interviewer	C6	C7	C8	C9	C10
group	C	H	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	C	C	H	C	C
Honest interviewer	H6	H7	H8	H9	H10
group	C	C	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 14 5가지 지표에 의한 K-means clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	C	H	C
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	C	C
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 15 부정행위를 한 조사원(Cheating interviewer)과 정상적으로 조사를 한 조사원(Honest interviewer) 탐색율(k-means clustering)

	기타 응답 비율	극단적 응답율	응답 거부율	Benford x^2	여과문항 응답율	5 가지 지표
Identified Cheating Interviewer (%)	100	50	20	60	60	90
Identified Honest Interviewer (%)	83.33	66.67	75	100	50	100

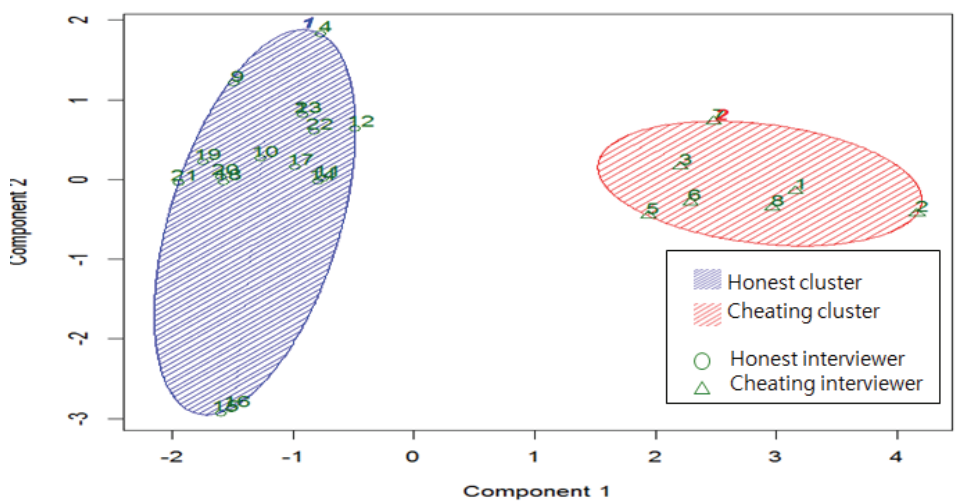

Figure 8 5가지 지표에 의한 k-means clustering (Honest cluster N=13, Cheating cluster N=9)

워드 방식(Ward method)의 클러스터링(clustering)은 각 개체를 하나의 클러스터로 정의하고 각 클러스터를 merge 하였을 때, 제곱 오류의 합(Error sum of squares)이 가장 작은 값이 나올 때까지 반복하여 구한다.

단계 1) 각 개체를 하나의 클러스터로 정의한다.

단계 2) 가장 가까운 클러스터를 merge 한다. 가장 가까운 클러스터는 merge한 후에 제곱 오류의 합(E)이 가장 작은 경우를 말한다.

$$E_k = \sum_{i=1}^{n_k} \left\| z_{ik} - \overline{z_k} \right\|^2$$

$$E = \sum_{k=1}^{c} E_k$$

단계 3) 원하는 클러스터의 수에 도달할 때까지 단계 2)를 반복한다.

Figure 9 ward clustering 알고리즘

<Table16>은 기타 응답 비율(other-answers-ratio)에 의한 ward clustering의 결과이며 <Table22>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 100%로 가장 높게 나왔으며, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 16.67%이다.

<Table17>은 극단적 응답율(extreme-answers-ratio)에 의한 ward clustering의 결과이며 <Table22>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 80%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 8.3%이며 극단적 응답율(extreme-answers-ratio)을 이용해서 정상적으로 조사를 한 조사원(honest interviewers)을 탐색하기 힘든 것을 알 수 있다.

<Table18>은 응답거부율(non-response-ratio)에 의한 ward clustering의 결과이며 <Table22>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 40%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약

33.33%라는 것을 알 수 있다.

<Table19>는 Benford x^2에 의한 ward clustering의 결과이며 <Table22>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 60%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약 100%로 가장 높은 것을 알 수 있다.

<Table20>은 여과 문항 응답율(filter-question-answers-ratio)에 의한 ward clustering의 결과이며 <Table22>를 통해 부정조사를 한 조사원(cheating interviewers)을 탐색한 비율이 60%, 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약 91.67%로 단변량 분석에서 가장 높은 것을 알 수 있다.

<Table21>은 기타 응답 비율(other-answers-ratio), 극단적 응답율(extreme-answers-ratio), 응답거부율(non-response-ratio), Benford x^2, 여과 문항 응답율(filter-question-answers-ratio)의 5가지 지표를 모두 포함한 다변량 분석(multivariate analysis)을 한 결과이다. C3와 C5를 제외하고 실제로 부정행위를 한 조사원(cheating interviewers)을 탐색한 비율이 60%이며 정상적으로 조사를 한 조사원(honest interviewers)을 탐색한 비율은 약 91.67%이다.

단변량 분석(univariate analysis) 중에서 k-means clustering과 마찬가지로 기타 응답 비율(other-answers-ratio)을 이용한 방법이 부정행위를 한 조사원(cheating interviewers)을 탐색한 비율이 높았고, Benford x^2은 정상적으로 조사를 한 조사원(honest interviewers)을 탐색하는 비율이 가장 높았다.

Table 16 기타 응답 비율(other-answers-ratio)에 의한 ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	C	C	C
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	C	C
Honest interviewer	H1	H2	H3	H4	H5
group	C	C	C	C	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	C	C	C	C
Honest interviewer	H11	H12			
group	C	C			

Table 17 극단적 응답율(extreme-answers-ratio)에 의한 ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	C	H
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	C	C
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	C	H			

Table 18 응답거부율(non-response-ratio)에 의한 ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	H	H
Cheating interviewer	C6	C7	C8	C9	C10
group	H	C	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	C	H	C
Honest interviewer	H6	H7	H8	H9	H10
group	H	C	C	C	C
Honest interviewer	H11	H12			
group	C	H			

Table 19 Benford x^2 에 의한 ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	C	H
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 20 여과 문항 응답율(filter-question-answers-ratio)에 의한 ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	H	C
Cheating interviewer	C6	C7	C8	C9	C10
group	C	H	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	C	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	H			

Table 21 5가지 지표에 의한 Ward clustering 결과

Cheating interviewer	C1	C2	C3	C4	C5
group	C	C	H	C	H
Cheating interviewer	C6	C7	C8	C9	C10
group	C	C	C	H	H
Honest interviewer	H1	H2	H3	H4	H5
group	H	H	H	H	H
Honest interviewer	H6	H7	H8	H9	H10
group	H	H	H	H	H
Honest interviewer	H11	H12			
group	H	C			

Table 22 부정행위를 한 조사원(Cheating interviewer)과 정상적으로 조사를 한 조사원(Honest interviewer) 탐색율(ward clustering)

	기타 응답 비율	극단적 응답율	응답 거부율	Benford x^2	여과문항 응답율	5 가지 지표
Identified Cheating Interviewer (%)	100	80	40	60	60	60
Identified Honest Interviewer (%)	16.67	8.33	33.33	100	91.67	91.67

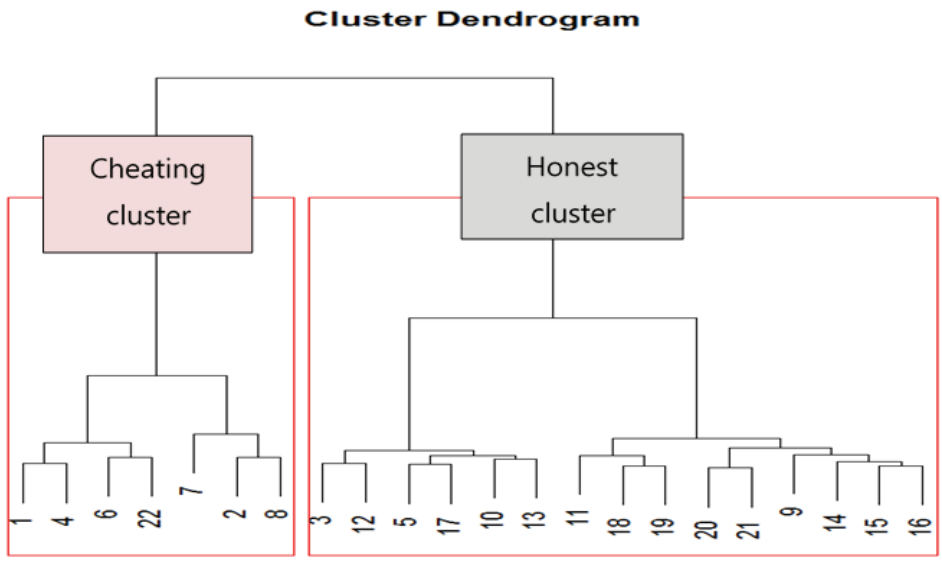

Figure 10 5가지 지표에 의한 ward clustering (Honest cluster N=15, Cheating cluster N=7)

3. 질 관리 지표 비교

본 조사와의 응답일치도를 평가하는 문항은 6개로 평생 음주 여부, 최근1년 음주 여부, 자동차 운전 여부, 최근 1년 스케일링 여부, 최근 1년간 독감예방접종 여부, 당뇨병 의사 진단 여부이다. 이러한 항목은 객관적 응답이 가능하며 보건학적 중요성을 고려하여 선정되었다.

<Table23>은 일치도 6개 문항 중 모두 일치하는 즉, 응답일치도가 6인 총 18명의 조사원의 표본가구대체율, 가구완료율, 응답일치도, 조사 소요시간 이상치 비율에 대한 표이며 <Table24>는 2012년 지역사회건강조사 폐기자료에서 부정행위 등으로 폐기된 조사원 65명의 질관리 지표를 나타낸 것이다.

2012년 지역사회건강조사 폐기자료에서 부정행위 등으로 폐기된 자료는 총 256건이며 응답일치도가 6인 조사 자료는 288건이다.

<Table23>과 <Table24>를 통해서 표본가구대체율의 평균은 일치도 6인 정상조사원(honest interviewer)와 부정행위 등으로 폐기된 자료의 조사원(cheating interviewer)의 차이가 없었고 부정행위를 한 조사원(cheating interviewer)의 가구완료율이 오히려 더 높았다. 문항일치도의 경우 예상과는 다르게 일치도가 6이라고 정의한 정상조사원(honest interviewer)과 일치도의 평균이 5.36인 부정행위를 한 조사원(cheating interviewer)과 크게 차이가 나지 않았다. 조사 건수 자체에서 부정행위를 한 조사원(cheating interviewer)이 더 적기 때문에 이와 같은 현상이 발생할 수도 있고 문항일치도 하나의 지표를 이용해서 조사원의 부정행위 유무를 판단하기 어려울 가능성도 존재한다.

반면에 조사 소요시간이 9분 이내로 조사 소요시간 이상치 비율을 비교했을 때 정상조사원(honest interviewer)와 부정행위 등으로 폐기된 자료의 조사원(cheating

interviewer)의 차이가 가장 컸으며 부정행위를 한 조사원의 조사 소요시간 이상치 비율 평균값이 약 0.47%이고 최대값이 11.34%로 정상조사원(honest interviewer)과 크게 차이가 난다.

Table 23 일치도 6인 정상조사원(honest interviewer)의 질 관리 지표

N=18	표본가구대체율 (%)	가구 완료율 (%)	문항일치도 (개)	조사 소요시간 이상치 비율 (%)
평균	7.5	92.09	6	0.08
표준편차	5.2	6.53	0	0.02
최대값	18.57	100	6	0.09
중위수	6.52	91.74	6	0
최소값	0	80	6	0

Table 24 부정행위를 한 조사원(cheating interviewer)의 질 관리 지표

N=65	표본가구대체율 (%)	가구 완료율 (%)	문항일치도 (개)	조사 소요시간 이상치 비율 (%)
평균	7.27	96.55	5.36	0.47
표준편차	6.67	3.17	0.37	2.13
최대값	34.57	100	6	11.34
중위수	5.06	98.48	5.4	0
최소값	0	86.21	4.31	0

Ⅳ. 결론 및 고찰

설문 조사는 부정행위를 하는 조사원(cheating interviewer)에 의해 잠재적으로 영향을 받게 되고 조사 중 부정행위는 심각한 bias를 야기시키는 것으로 밝혀졌다. 더욱이 부정행위를 하는 조사원(cheating interviewer)의 수가 적음에도 불구하고 통계결과에 심각한 문제를 초래한다. 현존하는 많은 방법들이 이러한 문제를 해결하기 위해 부정행위를 하는 조사원을 추출하기 위해 재조사(re-interview)를 시행한다. 재조사(re-interview)는 부정행위를 하는 조사원(cheating interviewer)를 찾는 가장 흔한 방법이지만 비용이 많이 드는 문제로 조사한 전체 대상자에 대해 재조사(re-interview)를 하는 것은 사실상 불가능하며 부정행위를 할 가능성이 있는 조사원(at risk interviewer)을 탐색하고 탐색한 조사원에 대해 재조사(re-interview)를 하는 것이 가장 이상적이다.

지역사회건강조사의 재조사(re-interview)는 전체 조사 대상자 약 22명의 10%를 랜덤으로 뽑아 전화점검을 실시하며 책임대학에서 부정행위가 의심되는 조사원을 전화점검 업체에 보고하여 그 조사원을 추가로 전화점검을 하여 조사 진행과정에서 부정행위 가능성을 확인한다. 전화점검을 통하여 6가지 문항 일치도 확인 만으로 부정행위 유무를 판단하는 것은 한계가 있다.

본 연구에서 기타 응답 비율(other-answers-ratio), 극단적 응답율(extreme-answers-ratio), 응답거부율(non-response-ratio), Benford x^2 , 여과 문항 응답율(filter-question-answers-ratio)의 5가지 지표를 이용하여 정상적으로 조사를 완료한 조사원(honest interviewer)과 부정행위를 한 조사원(cheating interviewer)을 비교하였지만 다섯 가지 지표들이 조사원들을 완전하게 구분 할 수 있는 기준이 되는 것은 아니었다.

이러한 결과가 나타날 수 밖에 없는 연구의 한계점은 다음과 같다. 첫째로, 부정

행위를 한 조사원의 수가 정상적으로 조사를 완료한 조사원(honest interviewer)의 수 보다 적었다. 이것은 조사과정에서 조사원이 부정행위(cheating)를 하는 경우는 크게 의도적인 행동과 비의도적인 행동으로 나뉘는데 적발된 데이터의 대다수가 비의도적, 예를 들면 귀화한 외국인인 줄 알았으나 알고 보니 귀화하지 않은 외국인이라 조사 대상자에서 제외가 되는 경우, 집 호수를 잘못 알아 다른 대상자를 방문하여 조사 자료가 폐기 된 경우가 대다수였고 이러한 경우는 본 연구에서 제외시켰기 때문에 부정조사를 한 조사원의 N수가 줄어들었다. 또한, 의도적인 행동으로 조사를 하였다고 할지라도 즉, 조사원이 조사 대상자가 없음에도 불구하고 대상자의 가족들에게 대리 응답으로 조사를 진행 할 경우 조사원 자신이 임의로 의도적인 설문조사를 작성하는 것이 아니라 가족들의 대답을 들으며 작성하는 것이기 때문에 정상적으로 조사를 완료한 조사원(honest interviewer)의 데이터와 크게 차이가 나지 않는 결과가 발생하였다. 둘째로, Benford 법칙을 이용한 분석에서 연속형 변수 선택의 제한이었다. 예를 들면, 성인의 키와 몸무게와 같은 변수는 첫 번째 숫자가 제한적으로 정해져 있고 지역사회건강조사의 변수 형태가 혈압 측정 횟수와 같은 최대값이 존재하는 변수가 대다수였기 때문에 변수 선택의 제한점이 있어 다양한 변수가 포함된 분석이 불가능 하였다.

하지만 본 연구를 통해서 발견한 사실은 다음과 같다. 다변량 분석(multivariate analysis)을 실시한 결과 부정조사원을 탐색할 수 있는 확률이 90%, 80% 이다. 전체 지역사회건강조사의 데이터에서 부정 행위를 한 조사원(cheating interviewer)과 부정행위를 하지 않고 정상적으로 조사를 끝낸 조사원(honest interviewer)을 구분하고 전체 데이터에서 부정행위를 한 조사원(cheating interviewer)을 탐색하기 위해서 이러한 다변량 분석을 이용하는 것은 불가능하다. 하지만 다섯 가지의 지표를 모두 포함한 다변량 분석을 이용해서 재조사(re-interview)를 위한 샘플링의 방법을 제공할 수 있다. 현재 전체 22만 명에서 랜덤으로 10%를 추출하여 재조사를 시행하고 있지만 앞에서 분석한 다섯 가지의 지표를 통해 부정행위를 할 가능성이 있는 조사

원(at risk interviewer)을 탐색 할 수 있다.

현재 전화 점검에서 지역사회건강조사 미참여, CAPI 미사용, 상품권 미수령과 함께 일치도 문항을 통해 일치도를 검증하고 있다. 부정행위를 한 조사원(cheating interviewer)와 일치도가 6인 정상 조사원(honest interviewer)의 질관리 지표 분석 결과 표본가구대체율, 가구완료율, 문항일치도에서는 큰 차이가 없었고 9분 이내의 조사를 한 조사 시간 이상치 비율이 부정행위를 한 조사원(cheating interviewer)에서 높게 나왔다. 2012년 지역사회건강조사의 281 문항을 노트북으로 기입하기 위해서는 최소 12~15분 정도의 시간이 필요한데 9분 이내에 설문 문항을 작성하였다는 것은 부정행위를 할 가능성이 높은 조사원(at risk interviewer)로 가정하여 심층 조사하여 부정행위를 탐색할 수 있으며 우선적으로 전화 조사를 이용한 재조사를 시행할 수 있다. 지역사회건강조사 질 향상을 위한 조사원 모니터링의 구체적인 제안 방법은 <Table25>에 제시하였다.

Table 25 부정행위를 할 가능성이 있는 조사원(at risk interviewer) 탐색 방법

		기존의 방법	제안 방법
전화 점검 이전	**표본 설계**	지역사회건강조사 자료업로드일 기준으로 보건소 별 10% 무작위 추출. 조사 수행 책임대학교에서 요청하는 전화점검대상자 추가 조사	지표 5가지를 이용한 다변량 분석을 통해 부정행위를 할 가능성이 있는 조사원(at risk interviewer)을 탐색하여 전화점검의 표본 추출
전화 점검 이후	**전화조사 후 피드백**	조사수행 책임대학교에서 조사 미참여, CAPI 미사용, 상품권 미수령에 해당하는 조사원 심층 조사 후 자료폐기 유무와 조사원 경고 및 해고 조치	조사소요시간이 정상적인 조사소요시간과 다르거나 시간대가 정상범위가 아닌 경우 부정행위를 할 가능성 있는 조사원(at risk interviewer)으로 가정하고 심층 조사

참 고 문 헌

김종희, 박해용, 김윤아, 강양화, 김정숙, 임도상, 탁양주, 임현술, 김호. 2008년 지역사회건강조사 질관리정도 평가. 한국보건정보통계학회지. 2010;35(2):87-100

질병관리본부. 2012년 지역사회건강조사 조사수행지침서. 2012

통계청, 통계품질관리 이렇게 합니다, 2006

신선옥, 한국노동패널조사의 응답자 태도에 면접원에 미치는 효과, 2008

Benford, F. (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society 78(1), 551–572.

Biemer, P. and S. Stokes (1989). The optimal design quality control sample to detect interviewer cheating. Journal of Official Statistics 5(1), 23–29.

Bushery, J., J. Reichert, K. Albright, and J. Rossiter (1999). Using date and time stamps to detect interviewer falsification. In Proceedings of the American

Statistical Association (Survey Research Methods Section), pp. 316–320..

Hill, T. (1999). The difficulty of faking data. Chance 26, 8–13.

Hood, C. and M. Bushery (1997). Getting more bang from the reinterviewer buck: Identifying ‘at risk’ interviewers. In Proceedings of the American Statistical

Association (Survey Research Methods Section), pp. 820–824.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4(1/4), 39–40.

Nigrini, M. (1996). A taxpayers compliance application of Benford’s law. Journal of the American Taxation Association 18, 72–91.

Nigrini, M. (1999). I’ve got your number. Journal of Accountancy 187(5),79–83.

Saville, A. (2006). Using Benford’s law to predict data error and fraud – an examination of companies listed on the JSE securities exchange. South African Journal of Economic and Management Sciences 9(3), 341–354.

Sch¨afer, C., J. Schr¨apler, K. M¨uller, and G. Wagner (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. Schmollers

Jahrbuch 125, 183–193. Schnell, R. (1991). Der Einfluss gef¨alschter Interviews auf Survey Ergebnisse. Zeitschrift f¨ur Soziologie 20(1), 25–35.

Schr¨apler, J. and G. Wagner (2003). Identification, charcteristics and impact of faked interviews in surveys - an analysis by means of genuine fakes in the raw data of SOEP. IZA Discussion Paper Series, 969.

Schreiner, I., K. Pennie, and J. Newbrough (1988). Interviewer falsification in census bureau surveys. In Proceedings of the American Statistical Association(Survey Research Methods Section), pp. 491–496.

Scott, P. and M. Fasli (2001). Benford's law: An empirical investigation and a novel explanation. CSM technical report, Department of Computer Science, University Essex.

Swanson, D., M. Cho, and J. Eltinge (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. In Proceedings of the American

Statistical Association (Survey Research Methods Section), pp. 4172–4177.

Bredl, S., N. Storfinger and N. Menold (2011). A literature review of methods to detect fabricated survey data. Discussion Paper 56. ZEU. Giessen.

Bredl, S., P. Winker and K. K¨otschau (2008). A statistical approach to detect cheating

interviewers. Discussion Paper 39. ZEU. Giessen.

Bushery, J.M., J. Reichert, K. Albright and J. Rossiter (1999). Using date and time stamps to detect interviewer falsification. In: Proceedings of the American Statistical Association (Survey Research Methods Section). pp. 316–320.

Chernick, M.R. (2008). Bootstrap Methods: A Guide for Practitioners and Researchers. Wiley. Hoboken, NJ. 2nd Ed.

Chipman, J.S. and P. Winker (2005). Optimal aggregation of linear time series models. Computational Statistics and Data Analysis **49**(2), 311–331.

Efron, B. (1978). Bootstrap methods: Another look at the Jackknife. The Annals of Statistics **7**(1), 1–26.

Efron, B. (1982). The Jackknife, the Bootstrap, and other Resampling Plans. Vol. 38 of CBMS-NSF Monographs. Society of Industrial and Applied Mathematics.

Forsman, G. and I. Schreiner (1991). The design and analysis of reinterview: An overview. In: Measurement Errors in Surveys (P.P. Biemer, R.M. Groves, L.E.

Lyberg, N.A. Mathiowetz and S. Sudman, Eds.). pp. 279–301. Wiley. Chichester.

Hauck, M. (1969). Is survey postcard verification effective?. Public Opinion Quar-

terly **33**, 117–120.

Hood, C.C. and M. Bushery (1997). Getting more bang from the reinterviewer buck: Identifying ‘at risk’ interviewers. In: Proceedings of the American Statistical Association (Survey Research Methods Section). pp. 820–824.

Jain, A.K. and J.V. Moreau (1987). Bootstrap technique in cluster analysis. Pattern Recognition **20**(5), 547–568.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. **16**(4), 277–298.

Abstract

Analysis and exploration of falsified interviewers in KCHS (Korean Community Health Survey)

Eunji Won

Dept. of Epidemiology and Biostatistics

Graduate School of Public Health

Seoul National University

Background

Korean Community Health Survey (KCHS), the nationwide community-based statistical survey for health, has been performed annually and performed for community health and various institutions and experts participate in this survey. Thus, quality control throughout the survey process is one of the most important parts for survey's accuracy and reliability. Data quality in face-to-face interviews like KCHS depends crucially by the interviewer, when he deviates from the prescribed interviewing procedure. If he does so consciously, this is referred to as interviewer falsification (Schreineret al., 1988) or cheating (Schrapler and Wagner, 2003). We identified characteristics of cheating interviewers unlike honest interviewer who didn't cheating interview and studied the effect of cheating interview data.

Aims

The goal of the current study is to compare cheating interviewers and honest interviewers and identify the characteristics of cheating interviewers. Also, we studied the identification of 'at risk' interviewers based on the above mentioned characteristics. This research will improve

KCHS's Quality Control, data's accuracy and reliability by identifying falsifications.

Methods

The data used in this study are KCHS's raw data in 2012, KCHS's discarded data in 2012, interviewer's Quality control Index data in 2012 and data of interviewer's characteristics in 2012.

As a novel tool in this context, cluster analysis are used and we us five indicator variables in the cluster analysis: the proportion of answers where the item 'others' including an alternative was selected in all answers which offered this item (referred to as others ratio), the proportion of 'extreme' ordinally scaled answers in all ordinally scaled answers referred to as extreme ratio, the proportion of answers where the item 'others' including an alternative was selected in all answers which offered this item (referred to as others ratio), the x^2-value stemming from the comparison of the leading digit distribution in all questionnaires of an interviewer with Benford's distribution, and the proportion of the filter question answered in the way that does not allow skipping part of the questionnaire.

Results

Cheating interviewers group (n=10) which had discarded data as proxy answers and honest interviewers group (n=12) without discarded data and cheating interview data is agreement of answer is 6. When compared to the five indicators in the two groups, the lower the value of the cheating interviewers' other-answers-ratio and extreme-answers-ratio, which was we supposed. Also, the higher value of cheating interviewers' Benford x^2 and filter-question-answers-ratio Index, which was we supposed. But non-answers-ratio and Benford x^2 values of the cheating interviewers' are higher than honest interviewer, which was we supposed differently.

We analyzed ward hierarchical cluster and k-means cluster using five indicators, the probability of exploring the cheating interviewer are 60% and 90%, respectively. In addition, when we compared total 65 interviewers of the discarded data (number of questionnaires: 256) and 18 interviewers who agreement of answer is 6 (number of questionnaires: 288), Index of outlier

time ratio (9minutes or less) is different of two groups.

Conclusion

Survey data are potentially affected by cheating interviewers and Interviewer cheating is a non-negligible problem as it can cause severe biases. Many of the existing methods dealing with the identification of fabricated interviews are derived from the survey design like re-interviews. The most common way to identify cheating interviewers is the re-interview but for reasons of expense, it is impossible to reinterview all households participating in a survey. So the question arises, how the reinterview sample can be optimized to best detect cheating interviewers. During KCHS survey period, we sampled 10% of survey subjects, performed a telephone survey and uploaded the data for efficient management. Based on the cluster analysis, five indicators are identified which might help to distinguish between cheating interviewer and honest interviewer. The result indicate that the indicators help to separate the two groups for reinterview sampling. And to identify the 'at risk' interviewers we consider Index of outlier time ratio.

Key words: Cheating Interviewer; Korean Community Health Survey; Quality Control; Benford's Law; cluster analysis

Student number: 2013-21869